36-710: Advanced Statistical Theory **Spring 2019**

Lecture 17: March 26

Lecturer: Alessandro Rinaldo *Scribes: Afra Feyza Akyürek*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

This lecture covers slow and fast rates for the Lasso with the restricted eigenvalue condition.

17.1 Slow rate for the Lasso

Last lecture we proved Theorem 17.1 which is known as slow rate for the Lasso.

Theorem 17.1 (Theorem 7.5 in the book.). *Let $\mathbf{X} \in \mathbb{R}^{nxd}$ and $\beta^* \in \mathbb{R}^d$ an unknown vector in the standard linear model $y = \mathbf{X}\beta^* + \epsilon$. If $\lambda = \lambda_n \geq \frac{\|\mathbf{X}^T\epsilon\|_\infty}{n}$, then any lasso solution $\hat{\beta}_{lasso}$ satisfies,*

$$MSE(\hat{\beta}_{lasso}) = \frac{\left\|\mathbf{X}(\hat{\beta}_{lasso} - \beta^*)\right\|^2}{n} \leq 4\,\|\beta^*\|_1\,\lambda_n.$$

Let's compare $MSE(\hat{\beta}_{lasso})$ to the $MSE(\hat{\beta}_{bss})$ for best subset selection.

$$MSE(\hat{\beta}_{bss}) \lesssim \|\beta^*\|_0\,\sigma^2\frac{log(ed/\delta)}{n} \qquad \text{with probability} \geq 1-\delta$$

$MSE(\hat{\beta}_{BSS})$ is almost optimal except the log term. Note that Theorem 17.1 is a deterministic result. In order to turn it into a more practical one, we ask the question: *When can we get w.h.p. that*

$$\lambda_n \geq \frac{\left\|\mathbf{X}^T\epsilon\right\|_\infty}{n} = \max_{i=1,\dots,n}\frac{|X_i^T\epsilon|}{n}$$

where X_i is the i'th column of $\mathbf{X}$? Assume $\max_i \|X_i\| \leq \sqrt{Cn}$ for some $C > 0$ (*), then,

$$
\begin{aligned}
P\left(\frac{\left\|\mathbf{X}^T\epsilon\right\|_\infty}{n} \geq \right) &= P\left(\max_{i=1,\dots,n} \frac{|X_i^T\epsilon|}{n} \geq t\right) \\
&= P\left(\max_{i=1,\dots,n} |X_i^T\epsilon| \geq tn\right) \\
&\leq \sum_{i=1}^{d} P\left(|X_i^T\epsilon| \geq tn\right) && \text{(Union Bound)} \\
&= \sum_{i=1}^{d} P\left(\frac{|X_i^T\epsilon|}{\|X_i\|} \geq \frac{tn}{\|X_i\|}\right) \\
& && \left(\tfrac{|X_i^T\epsilon|}{\|X_i\|} \in SG(\sigma^2) \text{ is unit vector}\right) \\
&\leq 2dexp\left(-\frac{t^2n}{2\sigma^2C}\right) \\
& && \text{(Using bound on SG variable and (*))}
\end{aligned}
$$

Then set RHS equal to δ and solve for t. Setting $t = \lambda_n = \sqrt{\frac{2\sigma^2C}{n}(log(2d) + log(1/\delta)}$ yields that $\lambda_n \geq \frac{\left\|\mathbf{X}^T\epsilon\right\|_\infty}{n}$ with probability $\geq 1-\delta$. Plug this into MSE for Lasso, then with probability $\geq 1-\delta$,

$$MSE(\hat{\beta}_{lasso}) \leq 4\left\|\beta^*\right\|_1 \sqrt{\frac{2\sigma^2C}{n}(log(2d) + log(1/\delta)}.$$

Note that this MSE still vanishes, but at a slower rate than BSS. It's off by a square-root. To further improve on $MSE(\hat{\beta}_{lasso})$, we put additional assumptions on the design matrix.

17.2 Fast rates for the Lasso

To obtain a faster rate of convergence we need stronger assumptions on $\frac{\mathbf{X}^TX}{n}$.

Definition 17.2 (Restricted Eigenvalue Condition ($Re(\alpha, \kappa, S)$))**.** *For $S \subseteq \{1,\dots,d\}$, $S \neq \emptyset$ and $\alpha \geq 1$, let*

$$C_\alpha(S) = \{\Delta \in \mathbb{R}^d : \left\|\Delta_{S^c}\right\|_1 \leq \alpha\left\|\Delta_S\right\|_1\} \qquad \textit{where} \qquad S^c = \{1,\dots,d\}\backslash S.$$

Then, we say that $X \in \mathbb{R}^{nxd}$ satisfies REC with respect to $S \subseteq \{1,\dots,d\}$ and parameters $\alpha \geq 1$ and $\kappa > 0$ if

$$\frac{1}{n}\left\|\mathbf{X}\Delta\right\|^2 \geq \kappa\left\|\Delta\right\|^2 \qquad \forall\Delta \in C_\alpha(S)$$

Intuition: Set $\Delta = \hat{\beta}_{lasso} - \beta^*$. Then $\frac{1}{n}\|\mathbf{X}\Delta\|^2$ is the MSE of $\hat{\beta}_{lasso}$. The function $z \to \frac{\|\mathbf{X}z\|^2}{n}$ can be very flat in the sense that $\frac{\|\mathbf{X}\Delta\|^2}{n}$ can be small but Δ can still be large.

If $\frac{\mathbf{X}^T\mathbf{X}}{n}$ has a minimum eigenvalue bounded away from zero, then

$$\|\Delta\|^2 \leq \frac{\frac{\|\mathbf{X}\Delta\|^2}{n}}{\lambda_{min}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)}$$

which requires that $\frac{\Delta^T\mathbf{X}^T\mathbf{X}\Delta}{n} \geq \lambda_{min} > 0$, $\forall \Delta \in \mathbb{R}^d$.

Theorem 17.3 (Theorem 7.2 in the book.)**.** *Assume that X satisfies REC w.r.t. $Re(\alpha, \kappa, S)$ where S is the support of β^*, then if $\lambda_n \geq 2\frac{\|\mathbf{X}^T\epsilon\|_\infty}{n}$, any lasso solution satisfies,*

$$MSE(\hat{\beta}_{lasso}) = \frac{\left\|\mathbf{X}(\hat{\beta}_{lasso} - \beta^*)\right\|^2}{n} \leq 9\lambda_n^2\frac{|S|}{\kappa}$$

where $9\lambda_n^2\frac{|S|}{\kappa} \simeq \frac{\|\beta^*\|_0}{\kappa}\frac{\sigma^2(log(d)+log(1/\delta))}{n}$ and

$$\left\|\hat{\beta}_{lasso} - \beta^*\right\| \leq \frac{3}{\kappa}\sqrt{|S|}\lambda_n$$

Also note that $|S| = \|\beta^*\|_0$.

Proof. We first need to show that $\Delta = \hat{\beta} - \beta^* \in C_3(S)$. By optimality of $\hat{\beta}$, it holds that

$$\frac{1}{2n}\left\|Y - \mathbf{X}\hat{\beta}\right\|^2 + \lambda_n\left\|\hat{\beta}\right\|_1 \leq \frac{1}{2n}\|Y - \mathbf{X}\beta^*\|^2 + \lambda_n\|\beta^*\|_1.$$

By rearraging, we obtain that,

$$\frac{1}{2n}\|\mathbf{X}\Delta\|^2 \leq \frac{\epsilon^T\mathbf{X}\Delta}{n} + \lambda_n(\|\beta^*\|_1 - \left\|\hat{\beta}\right\|_1).$$

Since β^* is S-sparse, i.e. $supp(\beta^*) = S$,

$$\begin{aligned}||\beta^*||_1 - ||\hat{\beta}||_1 &= ||\beta^*_S||_1 - ||\beta^*_S + \Delta_S||_1 - ||\hat{\beta}_{S^c}||_1 \\ &= ||\beta^*_S||_1 - ||\beta^*_S + \Delta_S||_1 - ||\Delta_{S^c}||_1\end{aligned}$$

(Since $||\hat{\beta}||_1 = ||\hat{\beta}_S||_1 + ||\hat{\beta}_{S^c}||_1$)

So, by Hölder's Inequality, it holds that,

$$\frac{1}{n}\left\|\mathbf{X}\Delta\right\|^2 \leq 2\frac{\left\|\mathbf{X}^t\epsilon\right\|_\infty}{n}\left\|\Delta\right\|_1 + 2\lambda_n$$

Now, using the facts that

1. $\left\|\beta_S^* + \Delta_S\right\|_1 \geq \left\|\Delta_S\right\|_1 - \left\|\Delta_S\right\|_1$
2. $2\frac{\left\|\mathbf{X}^T\epsilon\right\|_\infty}{n} \leq \lambda_n$.

$$\begin{aligned}\frac{1}{n}\left\|\mathbf{X}\Delta\right\|^2 &\leq [||\Delta_S||_1 + ||\Delta_{S^c}||_1 + 2||\Delta_S||_1 - 2||\Delta_{S^C}||_1] \\ &= \underbrace{\lambda_n}_{\Delta\in C_3(S)}\underbrace{(3||\Delta_S||_1 - ||\Delta_{S^c}||_1)}_{\geq 0}\end{aligned}$$

Next,

$$\begin{aligned}\lambda_n\left(3||\Delta_S||_1 - ||\Delta_{S^c}||_1\right) &\leq \lambda_n 3\left\|\Delta_S\right\|_1 \\ &\leq 3\lambda_n\sqrt{|S|}\left\|\Delta_S\right\| \\ &\leq 3\lambda\sqrt{|S|}\frac{\left\|\mathbf{X}\Delta_S\right\|}{\sqrt{n\kappa}} && \text{(by REC)}\end{aligned}$$

Thus,

$$\frac{1}{\sqrt{n}}\left\|\mathbf{X}\Delta\right\| \leq 3\lambda_n\sqrt{\frac{|S|}{\kappa}}$$

taking square of both sides gives

$$\frac{1}{n}\left\|\mathbf{X}\Delta\right\|^2 \leq 9\lambda_n^2\frac{|S|}{\kappa}$$

as claimed. Similarly, for the second part

$$\sqrt{\kappa}\left\|\Delta\right\| \underset{REC}{\leq} \frac{\left\|\mathbf{X}\Delta\right\|}{\sqrt{n}} \underset{\text{Above bound}}{\leq} 3\lambda_n\frac{\sqrt{|S|}}{\sqrt{\kappa}}$$

So, with probability $\geq 1-\delta$,

$$||\Delta|| = ||\hat{\beta} - \beta^*|| \leq 3\lambda_n\frac{\sqrt{|S|}}{\kappa} \lesssim \sqrt{\left\|\beta^*\right\|_\infty}\sigma\sqrt{\frac{(log(d) + log(1/\delta)}{\kappa}}$$

□